Supervised ML (Classification) on Company Bankruptcy Detection

Soma Pavan Kumar
Data science trainee,
AlmaBetter, Bangalore

Abstract:

Bankruptcy prediction is always a topical issue. The activities of all business entities are directly or indirectly affected by various external and internal factors that may influence a company in insolvency and lead to bankruptcy. It is important to find a suitable tool to assess the future development of any company in the market. The objective of this paper is to create a model for predicting potential bankruptcy of companies using suitable classification methods, namely Support Vector Machine and artificial neural networks, and to evaluate the results of the methods used.

Keywords: Data Wrangling, Pandas, Matplotlib, Seaborn, Machine Learning, Logistic Regression, K-Nearest Neighbors, Naive Bayes(Gaussian)

1.Problem Statement

Prediction of bankruptcy is a phenomenon of increasing interest to firms who stand to lose money because of unpaid debts. Since computers can store huge dataset pertaining to bankruptcy making accurate predictions from them beforehand is becoming important.

In this project you will use various classification algorithms on bankruptcy dataset to predict bankruptcies with satisfying accuracies long before the actual event

2. Introduction

In financial bankruptcy analysis, the diagnosis of companies at risk for bankruptcy is crucial in preparing to hedge against any financial damage the at-risk firms stand to inflict. A wide number of academic researchers from all over the world have been developing corporate bankruptcy prediction models, based on various modelling techniques. Numerous statistical methods have been developed. Despite the popularity of the classic statistical methods, significant problems relating to the application of these methods to corporate bankruptcy prediction remain.

The economic meltdown of 2008, initiated a conversation about market sustainability, and the tools that can be used to predict it. The need for better predictive models become apparent, in order to avoid such devastating events in the future.

Bankruptcy of companies and enterprises affects the financial market at multiple fronts, and hence the need to predict bankruptcy among companies by monitoring multiple variables takes on an added significance.

A better understanding of bankruptcy and the ability to predict it will impact affect the profitability of lending institutions worldwide

3. Data Preprocessing

Data preprocessing is a data mining technique which consists in transforming the data in order to make it understandable. It could be changing the type, the format, splitting the data, verifying that there are no missing values but also creating new columns thanks to columns we initially have. In machine learning, the data processing step is critical because it involves cleaning, integration, transformation, scaling, standardizing data and many other tasks, in order to have a good preparation for the application of models.

To begin we first did some data exploration by checking types, missing values and also duplicate values.

Pandas:

It has functions for analysing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Pandas allows us to analyse big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science. Pandas are also able to delete rows that are not relevant, or contain wrong values, like empty or NULL values. This is called cleaning the data.

Pandas library also provides us with functions such as describe () which provides us with statistics of the data including mean, mode, median and also Interquartile ranges.

Pandas library also gives us some functions like shape, which gives us the number of rows and columns in the whole data. Also, some functions like shape which gives us all the number of cells or elements in the whole of data.

Pandas also has some features like drop () which drops the whole column or a row as specified by the user.

4. Data Visualization

Data Visualization is an interdisciplinary field that deals with graphic representation of data. It is a particular efficient way of communication when the data is numerous as for example a time series

The first plot is maybe one of the most important because it shows all the correlations of the features. To complete this visualization, we created a ranking of the features which are the most correlated to the target. So, it gives an idea of which features we have to focus on.

Moreover, in the notebook, you can see that there are significant red squares in the plot indicating that there are significant high correlations.

All of the Data Visualizations In this project are done using Matplotlib.pyplot, which is an inbuilt library in python that deals with plotting graphs including, bar graphs, histogram, all the way till pie charts and scatter plots.

Matplotlib:

Matplotlib is a low-level graph plotting library in python that serves as a visualization utility. Matplotlib was created by John D. Hunter.

Matplotlib is open source and we can use it freely. Matplotlib is mostly written in python, a few segments are written in C, Objective-C and JavaScript for Platform compatibility.

Plotting in Matplotlib:

The plot () function is used to draw points (markers) in a diagram. By default, the plot () function draws a line from point to point.

The function takes parameters for specifying points in the diagram. Parameter 1 is an array containing the points on the **x-axis**. parameter 2 is an array containing the points on the **y-axis.**

In this project we've had 5 visualizations:

1. The First one is to find the Correlations between all the independent variables/Features.
2. The Second is to find the Distribution of the Target Variables.
3. The Third is to find the distributions of all the variables in a histogram
4. The fourth is to find some correlation between some variables using boxplots.
5. Also, plotting feature distribution for close to bankruptcy companies.
6. Also plotting the distributions of the variables after processing Log transformation on the variables.

Groupby Function:

A groupby operation involves some combination of splitting the object, applying a function, and combining the results. This can be used to group large amounts of data and compute operations on these groups.

Any groupby operation involves one of the following operations on the original object. They are −

- Splitting the Object
- Applying a function
- Combining the results

5. Evaluation metrics:

An evaluation metric quantifies the performance of a predictive model.

This typically involves training a model on a dataset, using the model to make predictions on a holdout dataset not used during training, then comparing the predictions to the expected values in the holdout dataset.

For classification problems, metrics involve comparing the expected class label to the predicted class label or interpreting the predicted probabilities for the class labels for the problem.

The evaluation metrics we used are:

1. **Accuracy**: Accuracy represents the number of correctly classified data instances over the total number of data instances

2. **Precision**: In an imbalanced classification problem with two classes, precision is calculated as the number of true positives divided by the total number of true positives and false positives.

3. **Recall:** In an imbalanced classification problem with two classes, recall is calculated as the number of true positives divided by the total number of true positives and false negatives.

4. **F-score:** F-Measure provides a way to combine both precision and recall into a single measure that captures both properties.

   Alone, neither precision or recall tells the whole story. We can have excellent precision with terrible recall, or alternately, terrible precision with excellent recall. F-measure provides a way to express both concerns with a single score.

5. **AUC-ROC Curve:** So, when it comes to a classification problem, we can count on an AUC - ROC Curve. When we need to check or visualize the performance of the multi-class classification problem, we use the AUC (Area Under the Curve) ROC (Receiver Operating Characteristics) curve. It is one of the most important evaluation metrics for checking any classification model's performance.

6. Data modelling:

We have a Classification problem because our target is whether a company is Bankrupt or not. So, the goal of this part is to apply many algorithms in order to find the algorithm with the best indicator. The indicators we decided to choose are the accuracy, precision, recall, F-score. This choice is because we wanted to be able to compare these algorithms between them and to choose which one is the most efficient. Let's apply Classification techniques to our problem.

6.1 Logistic Regression:

Logistic regression is a statistical model that in its basic form uses a logistic Function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

The AUC-ROC score is 0.84

6.2 K-Nearest Neighbors:

K nearest neighbors classifier is a non-parametric supervised learning algorithm first developed by Evelyn Fix and Joseph Hodges in 1951.

It works by finding the distance between a query and all the examples in the data. For classification problems, a class label is assigned on the basis of a majority vote—i.e. the label that is most frequently represented around a given data point is used.

To which example the query is closest to is calculated using different metrics such as Euclidean distance, Manhattan distance, Minkowski distance and other metrics.

The total accuracy score for KNN Classifier is 0.897, and a f1-score of 0.4 and an AUC-ROC score of 0.625

6.3 Naive Bayes(Gaussian) Classifier :

Naive Bayes classification is an algorithm that is based on the Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle.

The Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Naive Bayes classifier is also used for text classification algorithms.

The total accuracy score for XGBoost is 0.66 and f1-score is 0.16 and an AUC-ROC score of 0.76

7. Conclusion:

The Experiment I chose is to predict if a company is going bankrupt according to the given features.

The dataset has 65 features on offer.
We applied 4 Machine Learning Algorithms namely:

1. Logistic Regression
2. K-Nearest Neighbors
3. Naive Bayes (Gaussian)

According to the f1-scores and the Accuracy metrics, KNN is the best metric to choose for further predictions.
But, according to AOC-ROC, Logistic Regression stands first with a score of 0.84

Nevertheless, in this case, the best decision is to use Logistic regression because it can better recognize the minority class even misclassifying some not close to bankruptcy companies as close to bankruptcy.

Also, the parameters that contribute to bankruptcy of a company are also external such as decisions taken by the CEO of the company and dwindling workforce and so on. Which can neither be calculated nor predicted.

References:

- W3 schools
- Geek for geek
- Stack overflow
- Analytics Vidhya